DOCUMENT RESUME

ED 298 864 HE 021 855

AUTHOR Erwin, T. Dary
TITLE The Analysis of Ratings Using Generalizability Theory for Student Outcome Assessment. AIR 1988 Annual Forum Paper.
PUB DATE May 88
NOTE 13p.; Paper presented at the Annual Forum of the Association for Institutional Research (28th, Phoenix, AZ, May 15-18, 1988).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; *Generalizability Theory; Higher Education; Institutional Research; *Interrater Reliability; *Outcomes of Education; *Rating Scales; *Student Evaluation; Test Interpretation
IDENTIFIERS *AIR Forum; College Outcomes Assessment

ABSTRACT

Rating scales are a typical method for evaluating a student's performance in outcomes assessment. The analysis of the quality of information from rating scales poses special measurement problems when researchers work with faculty in their development. Generalizability measurement theory offers a set of techniques for estimating errors or unreliability due to a variety of errors. In this study, generalizability calculations were made for two sets of data, one from the arts and one from the sciences. Followup decision studies were calculated to determine the adequacy of the number of raters (interrater reliability) and the number of rating items. The rationale, terminology, and implications of generalizability theory are discussed for institutional researchers. Contains 7 references. (Author/KM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298864

The Analysis of Ratings Using Generalizability Theory For Student Outcome Assessment

T. Dary Erwin
Office of Student Assessment
James Madison University
Harrisonburg, VA 22807

Paper Presented at the Annual Meetings
Of the Association for Institutional Research
Phoenix, AZ

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

May 18, 1988

Running head: Generalizability and Assessment

HE 021 855

AIR

for Management Research, Policy Analysis, and Planning

This paper was presented at the Twenty-Eighth Annual Forum of the Association for Institutional Research held at the Hyatt Regency Phoenix Hotel, Phoenix, Arizona, May 15-18, 1988. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Teresa Karolewski
Chair and Editor
Forum Publications Editorial
Advisory Committee

Abstract

The application of generalizability measurement theory is a method for analyzing the assessment method of ratings for student outcome assessment. For examples, one from the arts and the other from the sciences are illustrated as typical problems an institutional researcher might face when implementing a program for outcome assessment. The rationale, the terminology, and the implications of generalizability theory are discussed for institutional researchers.

The Analysis of Ratings
Using Generalizability Theory
For Student Outcome Assessment

The clamor for educational reform came first, the call for accountability and assessment came second, and now comes the actual conducting of assessment activities. This role of assessment is often assigned to institutional researchers who are charged with performing outcome assessment studies or consulting with faculty in the design of such studies (Marchese, 1988). This study will illustrate the use of one type of assessment data, ratings, and the tools for their analyses called generalizability theory.

The measurement tools needed in assessment studies are sometimes new to the institutional researcher who has traditionally handled data not prone to reliability or measurement error such as student frequencies or square footage. Assessment data, on the other hand, contains error because it deals with human performance and is thus considerably less reliable.

This paper presents two examples, one from the arts and one from the sciences, of how to analyze ratings of students' work with generalizability measurement theory. A brief recap of the assessment movement will be discussed, then some common problems

institutional researchers will often confront in implementing an assessment program. Later, information will be analyzed from two outcome studies within an explanation of generalizability theory. These concepts and tools can then be applied by institutional researchers in a wide range of assessment projects.

The Charge for Assessment

The student outcome assessment movement is pervasive across the country in higher education today (Ewell,1987; Rossman and El-Khawas, 1987). Several states, essentially all regional accrediting associations, and many professional accrediting groups require colleges and universities to submit test score results and other assessment data for accountability purposes. About two-thirds of the states report they have formal initiatives for outcome assessment in place (Boyer, Ewell, Finney, and Mingle, 1987). Another survey (Rossman and El-Khawas, 1987) finds seven of every ten college or university administrators nationwide believe assessment should be linked to the institutional budget process.

Regardless of how assessment results are used, for student or program evaluation purposes, it is necessary that the assessment methods be sound and defensible. In practice, the quality of assessment methods or instruments can be judged in terms of its reliability and validity. Many institutions often choose a standardized test from a testing organization such as Educational Testing Service's Major Field Achievement Tests. In

those cases, most of the instruments from major testing organizations are reliable, and the institution should then determine the test's validity by matching it with a department's curriculum or program goals.

When a standardized instrument is inappropriate, for example, having low content validity, or an instrument does not exist for a particular program, the institution must design its own assessment method or instrument. Often, a multiple-choice test will be constructed; however, this type of instrument may not be desirable or appropriate for certain kinds of programs. Essays, projects, interviews, and artistic performances are other types of assessment methods faculty choose for end-of- program assessment.

With these types of assessment methods, rating scales are commonly designed to evaluate a student's performance. Ratings pose some unusual problems to institutional researchers because they contain at least one additional source of measurement error the subjectivity of the rater. Lack of agreement or low inter-rater reliability among raters, such as faculty, can yield unclear curriculum or program goals as well as different standards for student expectations. Basing decisions on assessment information with low reliability could lead to erroneous conclusions about the student's ability and quality of program.

Generalizability Theory

Generalizability measurement theory (Brennan, 1983) is one method for analyzing inconsistency of raters or for analyzing multiple sources of variation in a measurement. This approach may be contrasted with classical test theory which provides only a single, undifferentiated error component (Brennan & Kane, 1985). In most practical settings, it is more realistic to recognize that any measurement contains a variety of error components or sources of variation. Examples of sources of variation might include: 1) rater variance in assessing human performance, such as in a play, essay, or interview; 2) item variance such as test item scores, rating item values, or other types of measurement items; 3) person variance, or the fact that different persons are being tested. In generalizability terminology, these sources of variation are called facets, and the measurement itself is seen as a sample from a universe of admissible observations. A student's observed score then is decomposed into the universe score or rating, similar to the true score in classical test theory, and error scores or ratings corresponding to the several sources of variation. The process allows making appropriate generalizations about each facet.

The methodology for estimating the various error components is an adaptation of analysis of variance. Estimates of variance components are calculated for the facets and their interactions in a Generalizability study. In the follow-up Decision study,

the number of conditions for each facet may be adjusted to study the desired or optimum generalizability coefficient. The generalizability coefficient, interpreted like the reliability coefficient, is the ratio of the universe score variance to the observed score variance.

In this paper, two Generalizability studies with follow-up Decision studies were conducted. In the first example, a sample faculty (raters) rated the performance of exiting arts majors (persons) on a sample of ten performance areas (items). In the second example, another sample of faculty (raters) rated the science achievement of college seniors (persons) from three science problems (items). Sources of variance from each facet and an overall generalizability coefficient are reported and discussed below.

Results

Arts - Example. The overall generalizability coefficient for the 3 persons x 3 raters x 12 items design was .471. Since the generalizability coefficient ranges from 0 to 1 and may be interpreted similarly to a reliability coefficient, this value is too low for making decisions about this theatre program. The variance components were as follows: Persons = 0.001, Raters = 0.025, Items = 0.040, P x R = 0.053, P x I = 0.001, R x I = 0.000, and P x R x I = 0.512.

The persons x raters x items component has the greatest value, suggesting that increasing the number of raters and rating

items might improve the generalizability coefficient. Persons were fixed (versus random), because only three majors graduated. After constructing several Decision Studies (a method for adjusting the number of conditions of facets), a generalizability coefficient of .802 was reached with 5 raters and 22 items. This suggestion would then be given to the respective arts department: use two additional raters and add 12 additional rating items to the rating form.

Science Problems Example

The generalizability coefficient was .79 for the 7 persons x 3 raters x 3 items (problems). Variance components were: Persons = 13.86, Raters = 0.13, Items = 2.43, Persons x Raters = 0, Persons x Items = 9.82, Raters x Items = 0, Persons x Raters x Items = 2.07. Ideally, we would increase the number of persons because of the large variance component; however, it was unrealistic to expect more science majors from this particular program. Consequently, the number of items (problems) was increased to 4 in the Decision Study to arrive at a .80 generalizability coefficient. The problems test was recommended to be lengthened by one additional problem.

Physics - Oral Test

The generalizability coefficient was .71 for the 7 persons x 3 raters x 3 item design. Variance components were: Persons = 19.41, Raters = 189.11, items = 0, Persons x Raters = 17.19, Persons x Items = 0.39, Raters x Items = 77.90, Persons x Raters

x Items = 40.40. For this Decision Study, the number of raters was increased to 5 to arrive at a .82 generalizability coefficient. Based on this information, five raters were recommended for the science oral exam in the future.

Summary

Rating scales are a typical method for evaluating a student's performance in outcome assessment. The analysis of the quality of information from rating scales poses special measurement problems when researchers work with faculty in their development. Generalizability measurement theory offers a set of techniques for estimating errors or unreliability due to a variety of error components.

In this study, generalizability studies were calculated for two sets of data, an arts example and science example. Follow-up Decision studies were calculated to determine the adequacy of the number of raters (inter-rater reliability) and of the number of rating items. Both the number of raters and the number of rating items were adjusted to determine desired levels of overall reliability or generalizability.

Studies such as these will become more commonplace for the institutional researcher working in outcome assessment. The institutional researcher will be called upon to assist faculty and staff in the design and analyses of tests and rating scales for program evaluation. This paper presents one approach that can be used by institutional researchers in the implementation of

their new duties.

REFERENCES

Boyer, G. M., Ewell, P. T., Finney, J. E., and Mingle, J. R. (1987). Assessment and outcomes measurement: A view from the states. AAHE Bulletin, 8-12.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: The American College Testing Program.

Brennan, R. L. & Kane, M.T. (1985). Generalizability theory: Analyzing measurement error in the assessment of artistic products (ACT Technical Bulletin no. 49). Iowa City, IA: The American College Testing Program.

Ewell, P.T. (1987). Assessment: Where are we? The implications of new state mandates. Change, 23-28.

Marchese, T. J. (1988). Assessment, accreditation, and institutional effectiveness: Implications for our profession. In A.K. Dickey (Ed.), Managing Education Better: Technology and Tomorrow. Tallahassee: Association for Institutional Research.

Rossman, J.E., and El-Khawas, E. (1987). Thinking about assessment: Perspectives for presidents and chief academic officers. Washington, D.C.: American Council on Education.

Southern Association of Colleges and Schools (1988). Criteria for accreditation: Commission in colleges. Decatur, GA: Author.